BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Sistem Rekomendasi

Sistem rekomendasi merupakan suatu aplikasi untuk menyediakan dan merekomendasikan suatu item dalam membuat suatu keputusan yang diinginkan oleh pengguna (Ungkawa, et al., 2013). Penerapan rekomendasi didalam sebuah sistem biasanya melakukan prediksi suatu item, seperti rekomendasi film, musik, buku, berita dan lain sebagainya yang menarik user. Sistem ini berjalan dengan mengumpulkan data dari user secara langsung maupun tidak (Fadlil & Mahmudy, 2010).

Pengumpulkan data secara langsung dapat dilakukan sebagai berikut :

1. Meminta user untuk melakukan rating pada sebuah item.

2. Meminta user untuk melakukan rangking pada item favorit setidaknya memilih satu item favorit.

3. Memberikan beberapa pilihan item pada user dan memintanya memilih yang terbaik.

4. Meminta user untuk mendaftar item yang paling disukai atau item yang tidak disukainya.

Dalam pembangunan sistem rekomendasi, ada beberapa macam metode untuk menyelesaikan permasalahan, antara lain user-based collaborative filtering, content-based filtering, dan hybrid.

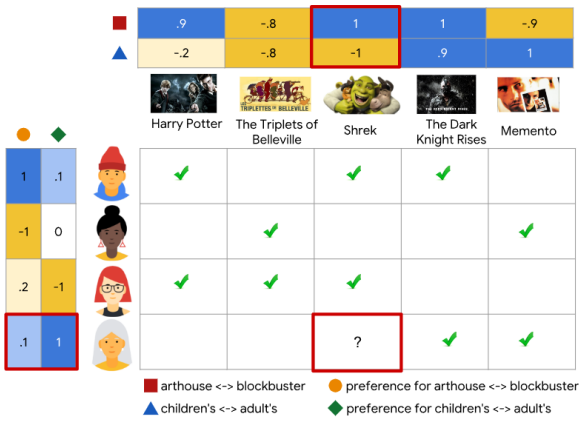
Namun beberapa peneliti menambahkan metode baru, yaitu knowledge based recommendation. Secara umum, Metode user-based collaborative filtering menggunakan feedback, ulasan dan rating untuk mendapatkan hasil rekomendasi. Metode content-based memberikan rekomendasi dengan membangun profile pengguna. Metode hybrid-based menggabungkan dua atau lebih metode.

Penggabungan dilakukan dengan tujuan saling melengkapi kekurang dari metode yang digunakan. Sedangkan metode knowledge-based menggunakan pola pengetahuan untuk memberikan hasil rekomendasi.

2.1.2 *Collaborative filtering* (CF)

Collaborative filtering (CF) merupakan proses penyaringan atau pengevaluasian item dengan menggunakan opini dari orang lain. Ide utamanya adalah untuk mengeksploitasi informasi mengenai perilaku di masa lampau maupun opini dari suatu komunitas pengguna yang kemudian digunakan untuk memprediksi item mana yang akan disukai atau menarik bagi seorang pengguna.

CF murni menggunakan matriks yang berisi user-item rating sebagai satu-satunya input, sedangkan output yang dihasilkan ada dua jenis: (1) prediksi (numerik) yang mengindikasikan seberapa besar tingkat kesukaan seorang pengguna terhadap sebuah item, dan (2) sebuah daftar berisi n item yang direkomendasikan. Istilah pengguna (user) dalam CF mengacu kepada mereka yang memberi penilaian terhadap item-item di dalam sistem, sekaligus nantinya menerima rekomendasi dari sistem (Dzumiroh, 2012). Berikut contoh Matriks preferensi collaborative filtering dapat dilihat di gambar 2.1 :



Gambar 2.1 Matriks preferensi collaborative filtering

2.1.3 *Missing Value*

Missing values dapat terjadi baik pada unit observasi maupun pada beberapa item pertanyaan saja (Handayani, 2011). Permasalahan *missing values* pada unit observasi dapat ditangani dengan menghapus *cases* yang mengandung m*issing value* kemudian melakukan modifikasi bobot dalam upaya penyesuaian untuk n*onresponse*. Menurut Little dan Rubin (2002), prosedur seperti ini dikenal dengan istilah *weighting procedures*.

Namun, untuk kasus *missing values* yang terjadi pada item pertanyaan, penanganan *weighting procedures* menjadi kurang efisien, hal ini dikarenakan *missing values* hanya terjadi pada beberapa item pertanyaan saja.

Menghapus item secara keseluruhan pada unit observasi,tentunya akan membuat hilangnya informasi yang telah dikumpulkan dan membuat pendugaan parameter menjadi tidak efisien. Menurut Little dan Rubin (2002), ketika *missing values* terjadi pada item pertanyaan, metode imputasi adalah prosedur yang dapat digunakan untuk menangani permasalah ini. Berikut ini adalah beberapa metode imputasi.

(1) Imputasi Metode mean

Mean merupakan salah satu metode imputasi yang paling umum digunakan. Imputasi dengan metode Mean mengisi missing data dalam suatu variabel dengan rata-rata dari semua nilai yang diketahui pada suatu variabel.

Imputasi dengan metode Mean memiliki kelemahan yaitu mengurangi varians pada variabel, karena nilai yang diisikan adalah sama untuk setiap variabel(Ilhamsyah.2015). Perhitungan awal dapat dilihat di persamaan (1) di bawah ini

$$pref_{u,g} = \frac{\sum_{i \in I_g} r_{ui}}{\|I_g\|} \quad (1)$$

$pref_{u,g}$ = nilai rating dari user u terhadap genre g,

r_{ui} = rating milik user u terhadap item i,

I_g = kumpulan item yang memiliki genre g.

Nilai imputasi mean yang akan digunakan pada pengisian rating sebuah item adalah nilai mean dari rating user tersebut terhadap genre-genre yang dimiliki item tersebut. Perhitungan imputasi mean terdapat pada persamaan (2) berikut.

$$impmean_{u,i} = \frac{\sum_{g \in G_i} pref_{u,g}}{\|G_i\|} \quad (2)$$

$impmean_{u,I}$ = nilai imputasi mean dari user u terhadap item i,

$pref_{u,g}$ = rating milik user u terhadap genre g,

G_i = kumpulan genre yang dimiliki item i.

Sebuah metode lain untuk imputasi juga digunakan, yaitu modus. Modus menunjukkan kecenderungan dari kumpulan data yang berbentuk diskrit. Pada dasarnya, bentuk rating pada Collaborative Filtering bermacam-macam. Untuk dataset film, data yang ada berbentuk numerik yang diskrit, sehingga metode modus juga dapat digunakan untuk imputasi. Terdapat nilai $impmode_{u,i}$ yang menunjukkan nilai imputasi modus untuk rating user u pada item i. Nilai tersebut adalah modus dari kumpulan rating terhadap item-item yang setidaknya memiliki satu buah kesamaan genre dengan item i.

Selanjutnya, nilai imputasi yang digunakan adalah perpaduan kedua imputasi, menggunakan sebuah parameter bobot berupa α dengan nilai 0-1. Parameter tersebut menunjukkan besarnya bobot imputasi mean, dan berbanding terbalik dengan bobot imputasi modus. Matriks rating yang padat (dense) akan terbentuk dengan

mengisikan nilai rating yang kosong menggunakan persamaan (3) dibawah ini :

$$impvalue_{u,} = \alpha * impmean_{u,} + (1- \alpha) * impmode_{u,i} \quad (3)$$

$imp\,value_{u,\mathrm{I}}$ = nilai imputasi terhadap rating user u untuk item i

$impmean_{u,\mathrm{I}}$ = nilai imputasi mean dari user u terhadap item i,

$imp\mathrm{mode}_{u,\mathrm{I}}$ = nilai imputasi modus dari user u terhadap item i.

(2) Imputasi Metode *Stochastic Hot Deck*

Hot-deck Imputation melibatkan penggantian missing values menggunakan nilai-nilai lain berdasarkan konsep similarity. Hot-deck Imputation salah satu metode imputasi yang popular digunakan. Meskipun popular dalam praktiknya, literatur tentang sifat-sifat teoretis dari berbagai metode sangat jarang.

Menurut Kowarik (2016), Hot-deck Imputation umumnya mengacu pada Sequential Hot-deck Imputation, yang berarti bahwa set data diurutkan dan nilai-nilai yang hilang diimputasikan secara berurutan berjalan melalui observasi demi observasi. Jika mengacu pada stokastik oleh gross (2008), proses stokastik adalah himpunan variable acak t$\{X(t), t \in T\}$. Semua kemungkinan nilai yang dapat terjadi pada variable acak X(t) disebut ruang keadaan(*state space*). Satu nilai t dari T disebut indeks atau parameter waktu. Dengan parameter waktu ini, proses stokastik dapat dibedakan menjadi dua bentuk yaitu :

A. Jika T = {0,1,2,3,…} maka proses stokastik ini berparameter diskrit dan biasanya disingkat dengan notasi $\{X_{12}\}$. (4)
B. Jika T = {t | t ≥ 0} maka proses stokastiknya berparameter kontinu dan dinyatakan dengan notasi {X(t) | t≥ 0}. (5)

2.1.4 RMSE (Root Mean Square Error)

RMSE adalah singkatan dari *Root Mean Square Error*, yaitu metrik evaluasi umum yang digunakan untuk mengukur kesalahan atau selisih antara nilai aktual dan nilai prediksi dalam suatu model atau analisis statistik. RMSE mengukur seberapa jauh rata-rata hasil prediksi dari nilai yang sebenarnya, dan digunakan untuk mengevaluasi performa model prediksi atau regresi, dan imputasi.

RMSE dihitung dengan cara menghitung selisih antara setiap nilai variabel prediktor dipilih berdasarkan hubungannya dengan variabel yang akan diimputasi kemudian men-squaring selisih tersebut, menjumlahkan kesalahan squared tersebut, dan kemudian mengambil akar kuadrat dari jumlah tersebut. (Grau, 2004). Lebih formalnya, Persamaan untuk RMSE dapat dilihat di persamaan (6) di bawah ini :

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_i)^2} \quad (6)$$

dimana:

$\hat{yi}$ = nilai prediksi observasi ke-i

yi = nilai aktual observasi ke-i

M = jumlah peramalan

Semakin kecil nilai RMSE, semakin baik performa model karena menunjukkan bahwa hasil prediksi model lebih dekat dengan nilai sebenarnya. Namun, RMSE harus selalu dibandingkan dengan nilai sebenarnya dalam suatu konteks tertentu, karena interpretasi nilai RMSE yang baik atau buruk sangat tergantung pada data dan domain aplikasinya.

2.1.5 MAE (*Mean Absolute Error*)

MAE (*Mean Absolute Error*) adalah salah satu metrik evaluasi yang digunakan untuk mengukur tingkat kesalahan atau selisih antara nilai aktual dan nilai prediksi dalam suatu model atau analisis statistik.

MAE mengukur seberapa jauh rata-rata hasil prediksi dari nilai sebenarnya. MAE dihitung dengan menjumlahkan selisih absolut antara setiap nilai aktual dan nilai prediksi, kemudian membagi jumlah tersebut dengan jumlah total observasi. Lebih formalnya, persamaan untuk MAE dapat dilihat di persamaan (7) dibawah ini :

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{7}$$

Dimana :

f_i adalah nilai hasil peramalan.

y_i adalah hasil sebenarnya.

n adalah jumlah data

Berdasarkan rumus 7 di atas, MAE menghitung rata – rata error dengan memberikan bobot yang sama untuk seluruh data (I = 1…n) secara intuitif. Untuk evaluasi model peramalan, MAE lebih intuitif dalam memberikan rata – rata error dari keseluruhan data. Dalam kasus ini pemilihan MAE menjadi tepat karena seluruh data diberikan bobot yang sama.

2.1.6 Basis Data

Basis data adalah kumpulan data yang disimpan secara sistematis di dalam komputer yang dapat diolah atau dimanipulasi menggunakan perangkat lunak (program aplikasi) untuk menghasilkan informasi.

Pendefinisian basis data meliputi spesifikasi berupa tipe data, struktur data dan juga batasan-batasan pada data yang kemudian disimpan. Basis data juga merupakan aspek yang sangat penting dalam sistem informasi karena berfungsi sebagai gudang penyimpanan data yang akan diolah lebih lanjut. Basis data menjadi penting karena dapat mengorganisasi data, menghindari duplikasi data, menghindari

hubungan antar data yang tidak jelas dan juga update yang rumit.(Zaenal Mustofa. 2021).

2.1.7 *Systematic Sampling*

Metode pengambilan sample acak sistematis adalah metode untuk mengambil sampel secara sistematis dengan interval (jarak) tertentu dari suatu kerangka sampel yangtelah diurutkan.

Menggunakan start point yang ditentukan secara judgement kemudian memilih tiap elemen populasi ke-n. Sampel dipilih berdasarkan interval yang ditentukan dari pembagian jumlah unit dalam populasi dengan jumlah sampel.

2.1.8 Google Colab

Google Colaboratory atau disingkat Google Colab adalah sebuah layanan dari Google yang menyediakan lingkungan pengembangan interaktif berbasis web (*web-based interactive development environment*) untuk penggunaannya dalam bidang pemrograman, terutama untuk pengembangan dan eksperimen di bidang Machine Learning dan Deep Learning.

Google Colab memungkinkan pengguna untuk membuat dan menjalankan notebook Python interaktif secara online, tanpa perlu menginstal perangkat lunak di komputer lokal. Notebook ini mencakup editor kode, sel pengkodean yang dapat dieksekusi, serta elemen tambahan seperti grafik, teks, gambar, dan catatan. Pengguna dapat mengakses notebook ini melalui browser web, dan juga dapat berkolaborasi dengan pengguna lain secara real-time.

2.2 Penelitian Terkait

Tabel 2.2 berikut merupakan daftar penelitian terkait yang menjadi rujukan atau refrensi pada penelitian ini :

Tabel 2.2 Penelitian Terkait

No	Nama Peneliti	Judul / Terbit	Algoritma	Uraian	Akurasi
1	Xiaoyuan Su, Taghi M. Khoshgoftaar	A Survey of Collaborative Filtering Techniques, 2009	Probabilistic memory-based collaborative filtering (PMCF)	Mengetahui Hasil tentang survey seberapa efektifnya *Collaborative filtering*(CF)	
2	Riyaz P A, Surekha Mariam Varghese	A Scalable Product Recommendations using Collaborative Filtering in Hadoop for Bigdata, 2015	Pearson correlation coefficient (PCC)	Hasil penelitian rekomendasi produk amazon menggunakan *collaborative filtering* dengan metode Hadoop	1 node = 80% 2 node = 65% 3 node = 52%
3	Fahmi Dhimas Irnawan, Indriana Hidayah, Lukito Edi Nugroho	Metode Imputasi pada Data Debit Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta, 2021	Multiple Imputation by Chained Equations (MICE)	Menggunakan metode imputasi dengan MICE dan k-NNi, pengisian data debit air sungai hilang	k-NNi = 80% MICE = 65%
4	Ikhsan Subagyo, Lukman	sentiment Analisis Review	SVM (Support Vector Machine)	pengujian klasifikasi	SVM = 87.620 %

	Dwi Yulianto, Wahyu Permadi, Arian Wahyu Dewantara, Anggit Dwi Hartanto	Film Di IMDB Menggunakan Algoritma SVM, 2019		menggunakan SVM dengan SGD memiliki nilai ketepatan yang hampir sama satu sama lain	SGD = 87.404 %
5	Azwar Rizal Alfarisi, Handayani Tjandrasa, dan Isye Arieshanti	Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma *Mutual Nearest Neighbor*, 2013	mutual nearest neighbour (MNN)	Imputasi dengan algoritma *mutual nearest neighbor* mempunyai performa yang lebih baik daripada imputasi dengan metode konvensional.	MNN = 91,76%
6	Iman7 Jihad Fadillah, Siti Muchlisoh	PERBANDINGAN METODE HOT-DECK IMPUTATION DAN METODE KNNI DALAM MENGATASI	Hot-deck Imputation	Metode Hot-Deck imputation lebih baik dibandingkan dengan metode k-NNi	k-NNi = 64,964% Hot-Deck = 83,626 %

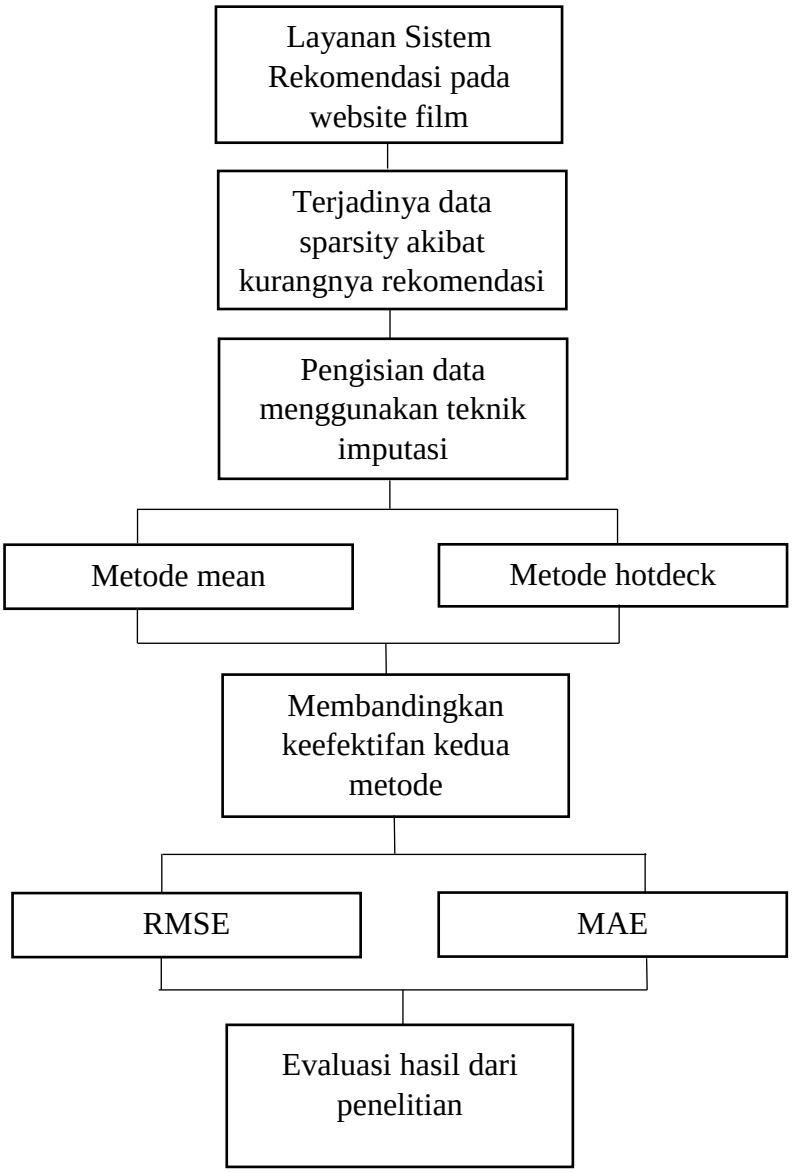
		MISSING VALUES, 2020			
7	Andreano, Christ and Lestari, Sri	IMPLEMENTASI ALGORITMA KNN, LOGISTIC REGRESSION DAN NAÏVE BAYES UNTUK KLASIFIKASI PENGAJUAN KREDIT PINJAMAN DI KOPERASI GENTIARAS PRINGSEWU LAMPUNG	KNN, Naive Bayes dan Logistic Regression	Algoritma Naïve bayes mendapatkan akurasi terbaik dibandingkan dengan algoritma KNN dan algoritma logistic regression.	Naïve bayes = 77.64% KNN = 76.64% Logistic Regression = 68.07%

Berdasarkan penelitian terdahulu , maka harus ada pebedaaan penelitan ini dengan penelitan terdahulu. Perbedaan dari penelitian yang diusulkan adalah:

1. Menggunakan dataset Movielens 100k.
2. Membandingkan metode Mean dan Metode Hot-Deck.
3. Pengerjaan menggunakan aplikasi web berbasi open-source Jupyter Notebook.
4. Evaluasi performa metric menggunakan RMSE (*Root Square Mean Error*)dan MAE (*Mean Square Error*).

2.3 Kerangka Pemikiran

Berdasarkan penjelasan teori diatas, maka selanjutnya peneliti akan menjelaskan alur kerangka penelitian. Untuk menemukan jawaban dari permasalahan data sparsity pada sistem rekomendasi film. Maka diperlukan teknik imputasi untuk mengisi kekosongan data, dan peneliti juga akan membandingkan antara kecepatan metode yang akan digunakan. Kerangka tersebut dapat dilihat di gambar 2.3 dibawah ini :



Gambar 2.3 Kerangka Pemikiran

2.4 Pertanyaan Penelitian

Pertanyaan yang dapat disimpulkan setelah mengetahui rumusan masalah di suatu penelitian. Rumusan masalah yang telah dinyatakan peneliti dalam bentuk kalimat pertanyaan. Dalam penulisan ini maka pertanyaannya sebagai berikut:

a. Bagaimana mengimputasi data yang hilang dari rating pengguna ?
b. Metode teknik pengimputasian apa yang efektif dalam pengimputasian ?